Data Intake Report

Name: G2M Insight for Cab Investment firm
Report date: 02/03/2021
Internship Batch:<Enter your batch code from Canvas course>
Version:<1.0>
Data intake by: Lauro Cesar Ribeiro
Data intake reviewer: Lauro Cesar Ribeiro
Data storage location: https://github.com/LauroCRibeiro/DataGlacier-Internship-DataAnalyst/tree/main/Week-2%20G2M%20insight%20for%20Cab%20Investment%20firm

Tabular data details: Cab_Data.csv

Total number of observations	359392
Total number of files	5
Total number of features	7
Base format of the file	.csv
Size of the data	21.8 MB

Tabular data details: City.csv

Total number of observations	20
Total number of files	5
Total number of features	3
Base format of the file	.csv
Size of the data	759 bytes

Tabular data details: Customer_ID.csv

Total number of observations	49171
Total number of files	5
Total number of features	4
Base format of the file	.csv
Size of the data	1.0 MB

Tabular data details: Transaction_ID.csv

Total number of observations	440098
Total number of files	5
Total number of features	3
Base format of the file	.csv
Size of the data	8.58 MB

Tabular data details: us_holidays.csv

Total number of observations	3288
Total number of files	5
Total number of features	10
Base format of the file	.csv
Size of the data	186 KB

Proposed Approach:

- I have found outliers in the Price_Charged feature, but I cannot see further information, such as that trip's duration. I decided not dealing with it as an outlier.
- I broke down the Date_of_Travel feature into other features. I would analyse more closely this way.
- All data were not missing values, so I just reshaped a bit the us_holidays.csv file to merge them properly.
- Ride profits are calculated by dividing the Price_Charged feature by Cost_of_Trip in each observation.
- Lots of categorical data.